DOCUMENT RESUME

ED 319 779 TM 015 000

AUTHOR Scheuneman, Janice Dowd
TITLE Assessing the Utility of Item Response Theory Models: Differential Item Functioning.
PUB DATE Apr 90
NOTE 13p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Boston, MA, April 17-19, 1990).
PUB TYPE Viewpoints (120) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Chi Square; Error of Measurement; *Estimation (Mathematics); Goodness of Fit; *Item Bias; *Item Response Theory; Mathematical Models; Maximum Likelihood Statistics; Methods Research; Research Problems
IDENTIFIERS Item Calibration; Item Characteristic Function

ABSTRACT

The current status of item response theory (IRT) is discussed. Several IRT methods exist for assessing whether an item is biased. Focus is on methods proposed by L. M. Rudner (1975), F. M. Lord (1977), D. Thissen et al. (1988) and R. L. Linn and D. Harnisch (1981). Rudner suggested a measure of the area lying between the two item characteristic curves (ICCs). Both signed and unsigned indices have been suggested, which may differ where the ICCs cross. Lord proposed asymptotic chi square tests for the equality of the "a" and "b" parameters as well as tests for each of these parameters separately. Thissen has more recently suggested a successive model fitting procedure where models to be tested can include those that constrain one or more of the three parameters for an item to be equal for two groups. The method developed by Linn and Harnisch calibrates the item only for the total group; the observed performance of the focal group is then compared to that expected from the IRT model. Issues raised by these methods are related to accuracy of item calibration, effects of estimation error, the problem of widely different ability distributions, goodness of fit, and problems associated with tests that are somewhat speeded or with data sets with many omits such as may occur on difficult tests under formula scoring instructions. Recent uses of parameter estimation methods based on marginal maximum likelihood methods using the BILOG program are discussed briefly. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED319779

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
JANICE SCHEUNEMAN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Assessing the Utility of Item Response Theory Models:

Differential Item Functioning

Janice Dowd Scheuneman

Educational Testing Service

In J. M. Ryan (Chair), Assessing the Utility of Item Response Theory Models. An Invited Symposium presented at meeting of the National Council on Measurement in Education, Boston, April 1990.

TM015000

BEST COPY AVAILABLE

Early in the 1970's the issues of what we then called item bias began to emerge as a problem that needed addressing. Over the next few years, a number of different procedures for detecting item bias were suggested and evaluated. Studies comparing these methods typically found that good results were obtained using a number of these procedures including those based on item response theory (IRT).

By around 1980, the IRT methods were generally perceived to be the methods of choice and a tendency arose to evaluate other procedures by how closely they approximated the IRT results. In fact, in their 1985 article Shepard, Camilli & Williams referred to a number of the non-IRT procedures as approximation methods. If it were not for the large sample size requirements and the cost of running the computer programs needed to estimate the item parameters (and assuming of course that the model fit the data), surely no other procedures would need apply.

Now it is 1990 and perhaps the time is right to take another look. At present, the perception remains that the IRT methods represent the ideal to which the other item bias methods aspire. In this paper, I would like to review these methods, consider some of the problems that have arisen and suggest perhaps that the IRT <u>models</u> represent an ideal toward which even the IRT <u>methods</u> can also only aspire.

Broadly, in the terms of IRT, an item is biased if its item characteristic curves (ICCs) differ for two groups of interest. In practice, however, several methods exist for assessing whether this is the case. This creates some ambiguity in the literature because, until fairly recently, comparisons of procedures have often cited the superiority of the IRT procedure, implying that the various methods are interchangeable.

In my presentation today I will focus on four of these methods with which I have some familiarity: Those proposed by Rudner (1975), Lord (1977), Thissen (Thissen, Steinberg, & Wainer, 1988) and Linn and Harnisch (1981). I am aware that there are others, or different indices derived from these, but I have chosen these four because I have used all of them with one or more data sets and can speak in part from experience.

Rudner (1975, 1980b) suggested a measure of the area lying between the two ICCs. Both signed and unsigned indices have been suggested, which may differ where the ICCs cross. Lord proposed asymptotic chi square tests for the equality of the *a* and *b* parameters, as well as tests for each of these parameters separately. Thissen (Thissen, Steinberg, & Wainer, 1988, 1989) has more recently suggested a successive model fitting procedure where models to be tested can include those that constrain one of more of the three parameters for an item to be equal for two groups. If such models fit the data as well as models in which they are permitted to differ, the item would be unbiased. The method suggested by Linn and Harnisch (1981) has been referred to as an approximation method because the item is calibrated only for the total group. The observed performance of the focal group is then compared to that expected

from the IRT model. This method is intended for use when the focal group is too small for separate calibrations.

Even in 1980 there were some caveats about the IRT methods for bias detection--the model didn't after all always fit the data and the programs for calibrating items required large samples, both of examinees and of items. The programs were also relatively expensive and difficult to run. (LOGIST, for example, is not user friendly.) Moreover, some of the strongest support for the IRT methods came from simulation studies where the data were generated from IRT models. From the perspective of several more years of research, one must question how well the assumptions made by these models are met or how robust the procedures are to violations of these assumptions.

First, can the items be calibrated with sufficient accuracy to make comparison of the ICCs for two groups meaningful? While it has long been recognized, for example, that the c parameters cannot be accurately estimated for many items, this problem has usually received only passing mention in the evaluation of IRT methods for bias detection. It is not at all unreasonable to suppose, however, that group differences exist in the likelihood of selecting the correct response in the absence of the relevant knowledge or ability. Such differences show up readily with some of the item bias methods conditioned on observed scores and have been observed in various analyses. Two problems exist then for the IRT methods: finding differences in the lower asymptotes if they exist and rejecting differences that are due only to estimation error.

The effects of estimation error may be largest for the area method. Particularly for relatively difficult items, the area between the two asymptotes can become quite large even if the two curves are otherwise very similar, depending somewhat on the bounds placed on the theta range.

In Lord's procedure, in recognition of the estimation problem, *c* parameters are estimated first on the entire group and then fixed for the separate estimation of the *a*'s and *b*'s. While this prevents identification of true differences in the *c*'s, it avoids the possibility of a serious error. It does not entirely solve the problem, however, because if the *c* has been fixed at a value that is a particularly poor fit, the effect may be to distort the estimate of the *a* parameter, possibly suggesting the presence of bias where it does not exist.

This problem might seem to be avoided with the Linn & Harnisch procedure where only one set of estimates are obtained. However, where mean scores for two groups are quite different, distributions at the low ability ranges may overlap minimally. Hence, the *c* parameter is likely to be estimated primarily on the performance of the lower scoring group, that is probably the focal group. Since it is the observed performance of the focal group that is compared to the curve obtained from the total, real differences between the groups in guessing ability are unlikely to be detected.

The Thissen method appears to me to deal most adequately with the problem of different *c*'s. If the data for either group is inadequate in the region of the asymptotes, the hypothesis of equal *c*'s will simply fail to be rejected. Different *c* parameters will be found only when sufficient the data

are available, making artifactual differences much less likely. Obviously, however, the method will fail to detect real differences in the same instances where it is a problem for the Linn and Harnisch method.

The problem of widely different ability distributions can effect more than the comparison of *c* parameters, however. In general, the ICC is best estimated in the region where most of the data from the calibration sample lies. In comparing ICCs of items calibrated on groups with different ability distributions (and possibly different sample sizes), the estimation errors will also differ quite possibly affecting the comparisons made to detect item bias. Methods that weight the area between the curve according to standard error have been suggested (Linn, Harnisch, Hastings, & Wardrop, 1981; Shepard, Camilli, & Williams, 1984.) Shepard, Camilli & Williams (1985), however, found that this statistic was less effective than the unweighted area measure when one of the groups became rather small.

More important, this difference between the ability distributions may affect the linking of scales which is needed for several of the IRT methods. The methods used in linking scales are essentially the same as some of those used for IRT equating, making that literature appropriate in this context as well. That is the next talk, so let me point out just one or two findings. First, Lawrence and Dorans (1990) found that slightly different equating results were obtained when samples from the two test forms differed in ability and when they were similar in ability. Cook, Eignor, and Taft (1988) also got different equating results on an achievement test using samples with different ability distributions. They explained these results in part, however, by

arguing that these samples were different also in terms of their academic preparation for this test and hence might be assumed to come from different populations. This is indeed a critical assumption for the IRT methods and one that might not be met in item bias applications. Can examinees from different population subgroups be assumed to come from the same more broadly defined test taking population? What exactly are we assuming when we use linking methods to place ICCs from different groups on scale?

A related question concerns model fit. While all agree that IRT methods cannot be used where the model does not provide a fit to the data, it seems clear that how robust the methods are to lack of fit differs with the application. For applications like bias detection, the focus is on individual items and comparisons are made of item characteristic curves derived from different samples differing perhaps by more than one standard deviation unit in mean ability. This application would appear to be more sensitive to misfit than equating, for example, where items are treated as a set and errors might be expected to average out over that set. The Linn and Harnisch procedure, where the observed performance in the reference group is compared to their performance from the model, may be particularly prone to artifactual indications of item bias if the item is misfit for the group as a whole.

Thissen's approach to bias detection appeals to me conceptually, but it has a serious drawback for practical applications. The analyses are based on response strings for all items in the analysis so that it does not work well for any but quite small numbers of items--probably fewer than 10. Thissen has suggested that this can be expanded by selecting a small set of four or five

anchor items, which are believed to be unbiased, and evaluating the other items by combining each separately with the anchor set. The selection of the anchor set, however, is at least as questionable as any assumptions made in linking items for separate calibrations. This also makes this procedure even more complex and time consuming that the more standard IRT procedures. Moreover, one of the criteria that has been used in contrasting different item bias methods is how easy they are to explain to a lay audience. Well, if you thought IRT methods generally were difficult to explain, wait until you try this one.

Finally IRT procedures in general do not deal well with tests that are somewhat speeded or with data sets with many omits such as may occur on difficult tests under formula scoring instructions. Ironson & Subkoviak (1979), for example, found that the IRT methods performed poorly on one of the tests they examined that appeared to be quite speeded. The situation becomes even more problematic if there may be differential speededness for the groups being compared.

Progress has been made in improving the IRT methods over the past few years. Some believe that the parameter estimation methods based on marginal maximum likelihood methods, such as are used by the BILOG program, may not suffer from some of the problems encountered by LOGIST. A recent study by Lim and Drasgow (1990) compares results obtained using these estimation procedures, Lord's procedure, and a Bayes modal estimation method. Other refinements to the methods, such as the weighted indices suggested by Linn et al (1981) for the area measure, continue to appear in the literature. A

recent review by Hills (1990) summarizes some of this newer work. Moreover, some problems in the application of bias detection models may be addressed best or perhaps only with IRT methods such as in testing situations where different examinees receive different subsets of items.

To conclude, let me state my own bias. Throughout this paper I have referred to the methods as item bias detection methods rather than differential item functioning (DIF) methods. I have no problem with the use of the DIF as terminology; I generally use it myself. I may have been among the first to point out that the item bias methods may detect any systematic difference between the performance of two groups, whether of not these differences might be thought of as bias. However, I also think it is important to remember the reason that these methods were developed. DIF may be what we find with our statistical procedures, but what we want to know about is item bias.

I believe that, over the past several years, we have devoted considerable talent and creativity in devising ever more refined and sophisticated methods to detect a very fuzzy and conceptually ill-defined phenomena. It is as if we were trying to use a tape measure, like that used by tailors, to measure the girth of some large furry animal--perhaps, a bear. When we find ourselves unable to do so very reliably, we go seek a tape calibrated in sixteenths of an inch instead of eighths. The problem, however, is not with our measure, but in the nature of the beast. We need to measure only within an inch or two in order to derive some notion of what we are dealing with.

At present, the most generalizable finding from item bias/DIF studies is that researchers find it difficult to interpret their results. Before significant progress can be made in this direction, we need solid, research-based theory concerning group differences in characteristic modes of learning and cognitive processing as well as theory on how these differences interact with test materials. We need to know much more about the properties of test items that make one item harder than another within one group before we can progress very far in understanding how an item might be differentially difficult for two groups.

I believe that research concerning differential group and item functioning may be one of the strongest approaches we have to help us toward this understanding, but I am unsure, given the estimation problems, that the IRT methods are indeed to be preferred to many of the simpler methods conditioned on observed score.

Bibliography

Bleistein, C. A. (1986). Application of item response theory to the study of differential item characteristics: A review of the literature (RM-86-3). Princeton, NJ: Educational Testing Service.

Burrill, L. E. (1982). Comparative studies of item bias methods. In R. A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore: Johns Hopkins University Press.

Candell, G. L., & Drasgow,F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. Applied Psychological Measurement, 12, 253-260.

Cook, L.L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. Journal of Educational Measurement, 25, 31-45.

Devine, P. J., & Raju, N. S. (1982). Extent of Overlap Among Four Item Bias Methods. Educational and Psychological Measurement, 42, 1049-1066.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. Applied Measurement in Education, 2, 313-334.

Hills, J. R. (1989). Screening for potentially biased items in testing programs. Educational Measurement: Issues and Practice, 8(4), 5-11.

Ironson, G. H. Use of chi-square and latent trait approaches for detecting item bias. In R. A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore: Johns Hopkins University Press.

Ironson, G. H., & Subkoviak, M. J. (1979). A comparison of several methods of assessing item bias. Journal of Educational Measurement, 16, 209-225.

Lawrence, I. M., & Dorans, N. J. (1990). Effect on equating results of matching samples on an anchor test. Applied Measurement in Education, 3, 19-36.

Lim, R. G., & Dragow, F. (1990). Evaluation of two methods for estimating item reponse theory parameters when assessing differential item functioning. Journal of Applied Psychology, 75, 164-174.

Linn, R. L., & Harnish, D. (1981). Interactions between item content and group measurement on achievement test items. Journal of Educational Measurement, 18, 109-118.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. Applied Psychological Measurement, 5, 159-173.

Lord, F. M. (1977). A study of bias using item characteric curve theory. In N. H. Poortinga (Ed.), Basic problems in cross-cultural psychology. Amsterdam: Swits & Vitlinger.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Rudner, L. M. (1977). An evaluation of select approaches for biased item identification. Unpublished doctoral dissertation, Catholic University.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980a). Biased item detection techniques. Journal of Educational Statistics, 5, 213-233.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980b). A Monte Carlo investigation of seven biased item detection techniques. Journal of Educational Measurement, 17, 1-10.

Scheuneman, J. D., & Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. Applied Measurement in Education, 1989, 2, 255-275.

Shepard, L. A., Camilli, G. & Averill, M. (1981). Comparison of procedures for detecting test-item bias using both internal and external ability criteria. Journal of Educational Statistics, 6, 317-375.

Shepard, L. A., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. Journal of Educational Statistics, 9, 93-128.

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22, 77-105.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In W. Wainer & H. I. Braun (Eds.), Test validity. Hillside, NJ: Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1989, October). Detection of Differential item functioning using the parameters of item response models. Paper presented at the Educational Testing Service conference on Differential Item Functioning, Princeton, NJ.